Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Vocal tract normalization in articulatory space using thin-plate spline method

Jianguo Wei[a] and Jianwu Dang[b]

[a]LTCI/CNRS, TSI/ENST, DB407, 37/39, rue Dareau, 75014 Paris, France
[b]Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, 923-1292 Ishikawa, Japan
wei@enst.fr

Inter-subject normalization is a key issue of group analysis of articulatory data to obtain a general description of kinematic properties of human speech production. Multi-subject articulatory study however is scarce due to the difficulty of normalization in articulatory domain. In order to reduce inter-subject variations among articulatory space, a simple normalization procedure was proposed using a Thin-plate splines method. The purpose of this normalization processing is to reduce the morphological differences of vocal tracts such as shape and size among different subjects. The Electromagnetic Articulographic (EMMA) data were used in our experiments, which were obtained from the NTT EMMA database for three subjects data included. A physiological articulatory model has been used to serve as the template. The landmarks were defined consistently in vocal tract space over the template and all subjects. The evaluations showed that the variances over subjects have been reduced 2.1mm for consonants and 2.3mm for vowels averaged over all tongue pellets.

1 Introduction

Inter-subject normalization for the data is a necessary processing step for group analysis of kinematical properties crossing age, gender, dialect and language. Electromagnetic Midsagittal Articulographic (EMMA) provides a good way to record the speaker's kinematical and acoustic information simultaneously, and some EMMA databases have been constructed. However most the analysis of EMMA data are based on single subject instead of multi-subject articulatory studies. One of the reasons of the scarcity of the multi-subject articulatory study is the difficulties of normalization. Bechman *et al.*[1] straighten the vocal tract wall to transform the coordinates for MRI data. Hashi *et al.*[2] normalized the vowel posture from x-ray microbeam database. These two methods both straighten the palate wall, which could not guarantee the relationship between palate and tongue after transformation, especially along the horizontal movement. In addition, these methods can not guarantee the constrained places of articulation after normalization, because they only linearly extended the palate wall, which could not reflect the nonlinear relationship between different subjects.

This research proposed a method to normalize the inter-subjects' EMMA data using Thin-plate splines transformation method that reflects a nonlinear relationship between subjects, which can keep the constraint places during transformation so as to overcome the shortage of the previous methods. The purpose of this research is to reduce the morphological variability of vocal tract shape and size over different subjects. This would be expected to facilitate the analyst to compute a single kinematical representation of entire group of subjects or to compare the kinematical representation between two different groups of subjects. Thin plate spline(TPS) warp [3] is a kind of radial based function which was a sort of widely applied transformation function in image alignment and shape matching. We adopted TPS to realize the Point-based normalization of different subjects' EMMA data in which a physiological articulatory model served as a template. The gridline system was considered to define the landmarks. Three subjects' EMMA data sets of NTT EMMA database[4] were used in the analysis. The evaluation show that the cross subject variability was reduced 2.1 mm in x-dimension and 2.3 in y-dimension by means of this procedure.

2 Thin Plate Splines Warping

Thin plate splines are a class of non-rigid spline mapping functions with several desirable properties for our application. They are globally smooth, easily computable, separable into affine and non-affine components, and contain the least possible non-affine warping component to achieve the mapping. Given a set of n corresponding 2D points, the TPS warp is described by 2(n+3) parameters which include 6 global affine motion parameters and 2n coefficients for correspondences of the control points. These parameters are computed by solving a linear system [5]. Suppose $(\hat{x}_i, \hat{y}_i) \in \Re^2$,i=1,…n, are the n control points in a planar, and their corresponding function values are, $\hat{v}_i \in \Re$, i=1,2,…,n, then the thin plate spline interpolation $f(x,y)$ denote a mapping: $f:\Re^2 \to \Re$. The TPS interpolating the points is defined by

$$f(x,y) = a_1 + a_2x + a_3y + \sum_{i=1}^{n} w_i r_i^2 \ln r_i^2 \qquad (1)$$

where $r_i^2 = (x-\hat{x}_i)^2 + (y-\hat{y}_i)^2$. Eq. (1) is the equation of a plate of infinite extent deforming under loads centred at $(\hat{x}_i, \hat{y}_i)$ .The plate deflects under the imposition of loads to take values w_i [5]. The interpolation spline function consists of two parts: affine transformation specified by former 3 elements, and the last warping part. The function f minimizes the bending energy E_f over the class of such interpolations where E_f is defined as:

$$E_f = \iint_{\Re} \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right) dxdy \qquad (2)$$

Three more equations are obtained using the following three constraints:

$$\sum_{i=1}^{n} w_i = 0 \qquad (3)$$

$$\sum_{i=1}^{n} \hat{x}_i w_i = 0 \qquad (4)$$

$$\sum_{i=1}^{n} \hat{y}_i w_i = 0 \qquad (5)$$

Constraint (3) shows that the sum of the loads applied to the plate should be 0. This is needed to ensure that the plate would not move under the imposition of the loads and remain stationary. Constraint (4) and (5) require that moments with respect to x and y axes are zero, ensuring that the plate would not rotate under the imposition of the loads.

The TPS parameter vectors ***a*** including a_1, a_2 and a_3 , and ***w*** including w_i ,can be computed by solving the following linear equation:

$$\begin{bmatrix} A & P \\ P^T & O \end{bmatrix} \begin{bmatrix} w \\ a \end{bmatrix} = \begin{bmatrix} v \\ 0 \end{bmatrix} \tag{6}$$

Where $A_{ij} = r_{ij}^2 \ln r_{ij}^2$, i=1,…n (the number of landmarks), j=1,…m (the number of raw data to be transformed); the i-th row of $\boldsymbol{P}$ is $(1, \hat{x}_i, \hat{y}_i)$. $\boldsymbol{O}$ is 3×3 matrix of zeros. The **0** is a 3 zero vector in the rightmost part of equation 6. ***w***, ***a*** and ***v*** are vectors formed from w_i, a_1, a_2, a_3 and v_i, respectively. The leftmost (n+3)×(n+3) matrix are denoted as ***K*** hereafter.

In this research, we focus on mapping points (x, y) of EMMA data to template coordinate (x', y') in light of given landmarks $(\hat{x}_i, \hat{y}_i)$ for one subject's EMMA data vs. $(\hat{x}_i', \hat{y}_i')$ defined for the landmarks of template. So we are interested in warping 2D points using TPS defined by pairs of control points. Toward that end, we applied TPS functions on *x* and *y* coordinates separately. From Equation 6, the TPS warp which maps $(\hat{x}_i, \hat{y}_i)$ to $(\hat{x}_i', \hat{y}_i')$, can be recovered by

$$\begin{bmatrix} w_x & w_y \\ a_x & a_y \end{bmatrix} = K^{-1} \begin{bmatrix} \hat{x}' & \hat{y}' \\ 0 & 0 \end{bmatrix} \tag{7}$$

Where $\hat{x}'$ and $\hat{y}'$ are the vectors formed with $\hat{x}_i'$ and $\hat{y}_i'$ respectively. The $\boldsymbol{w_x}$ and $\boldsymbol{a_x}$ are the parameters for x-dimension as well as $\boldsymbol{w_y}$ and $\boldsymbol{a_y}$ are the parameters for y-dimension, which have the same meanings as in equation 6. The transformed coordinates (x_j', y_j') of points (x_j, y_j) are given by

$$\begin{bmatrix} x' & y' \end{bmatrix} = [B \quad Q] \begin{bmatrix} w_x & w_y \\ a_x & a_y \end{bmatrix} \tag{8}$$

Where $B_{ji} = ((x_j - \hat{x}_i)^2 + (y_j - \hat{y}_i)^2) \ln((x_j - \hat{x}_i)^2 + (y_j - \hat{y}_i)^2)$, i=1,…,n, j=1,…,m. The *j*-th row of $\boldsymbol{Q}$ is $(1, x_j, y_j)$, and j-th row of the resulting vectors *x'* and *y'* are the interpolated x and y coordinates x_j' and y_j' ,respectively [6].

3 Landmark selection

In this research, we used a physiological articulatory model's vocal tract [7] as the template in the normalization process. The landmarks are firstly defined in the template, and then the corresponding landmarks are defined for EMMA data of each subject. There is no explicit way to define the identifiable feature points between the template and each subject along the palate, the most identified feature points are the points with respect to the coils (The T1 to T4) putting on the tongue surface. Therefore, we firstly calculated the average tongue positions along tongue surface (from tongue tip to tongue rear) over all the vowels in the database. And then we calculated the centroid point of the tongue surface based on the averaged tongue tip, tongue dorsum and tongue rear. After that, we link the centroid point to the tongue tip and tongue dorsum and extended to intersect with the palate. Next, we partition the angle between the tongue tip and tongue rear to eight equal fan sections, since the distance between the tongue rear and the tongue tip is roughly equal to eight centimetres along the tongue surface of the template. In order to cover the space after the tongue rear, we extended two more fan sections following the tongue rear. Totally, there are 10 equal fan sections along the vocal tract. We also interpolated the tongue surface based on the four points (T1-T4) by means of cubic spline. The middle line between the palate and tongue surface were also calculated. One additional line below the tongue surface was also calculated which has the three forth distance of tongue surface to the centroid point. Consequently, ten sub-fans' edges intersected with the palate line, the middle line, the tongue surface and the line below the tongue surface, so as to obtain 44 intersection points totally, which serve as the landmarks in the normalization. The landmarks of each subject were defined under the same procedure. The results were shown in the figure 1 for template and figure 2-4 for 3 subjects.

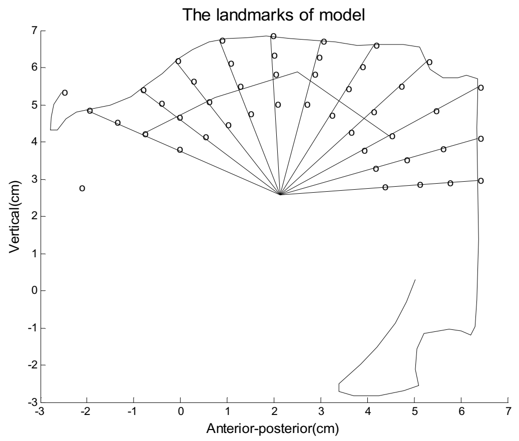

Figure 1. The figure shows landmarks of the template. The leftmost side is the lip side. The palate curve, the line linking the T1 to T4 and the grid line were drawn in this figure for the template. The circles are the landmarks.

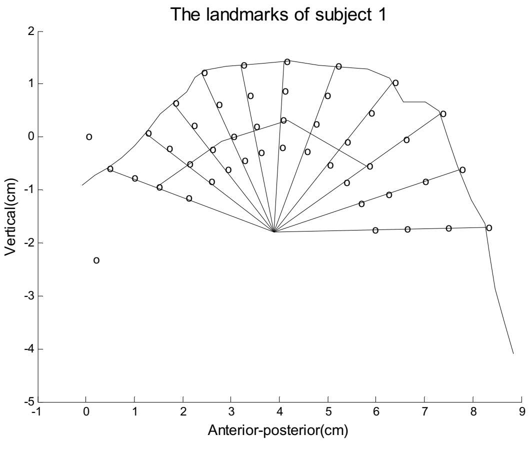

Figure 2. The landmarks of subject 1. The circles are the landmarks.

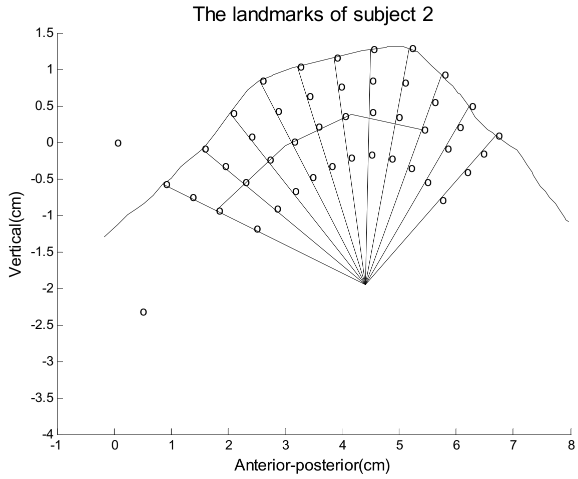

Figure 3. The landmarks of subject 2. The circles are the landmarks.

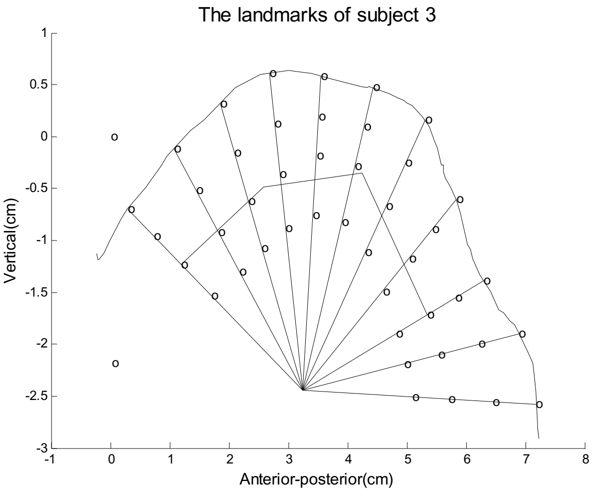

Figure 4. The landmarks of subject 3. The circles are the landmarks.

4 Experiments

In order to conduct the experiments, we extracted the distributions of 5 Japanese vowels and 8 consonants from the EMMA data. The steady states of vowels were recorded and the original distribution was shown as figure 5. And the stead states of consonants were shown in the figure 7. The solid curve is the palate curve of the template. While the dash-dot curves, dashed curve, and dotted curve are the palate curves of subject 1, 2 and 3 respectively.

Comparing the figure 5 with 6 and figure 7 with 8, one can intuitively find that the variances between different subjects were reduced. The palate curves of each subject were almost overlapped with the palate of the template.

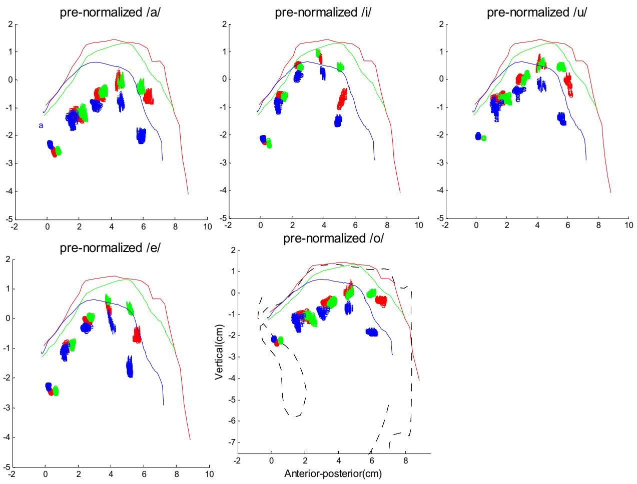

Figure 5. The raw data before normalization, each panel shows the data of one vowel of 3 subjects denoted by different color. The stars are tongue tip, cross symbols denote the tongue blade, triangles stand for tongue dorsum and circles depict tongue rears.

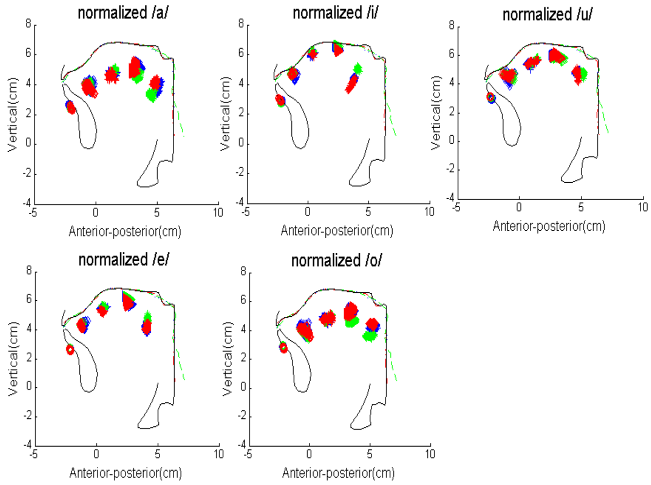

Figure 6. The distributions after normalization. The symbols have same meaning as figure 5.

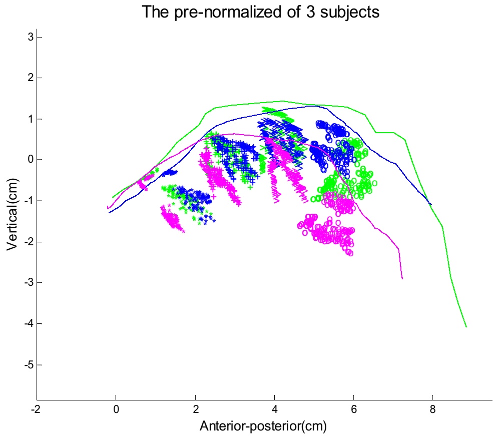

Figure 7. The non-normalized distributions of all consonants of three subjects are plotted. The left side is the lip. The 3 subjects are denoted by different colours. The stars are tongue tip, cross symbols denote the tongue blade, triangles stand for tongue dorsum and circles depict tongue rears.

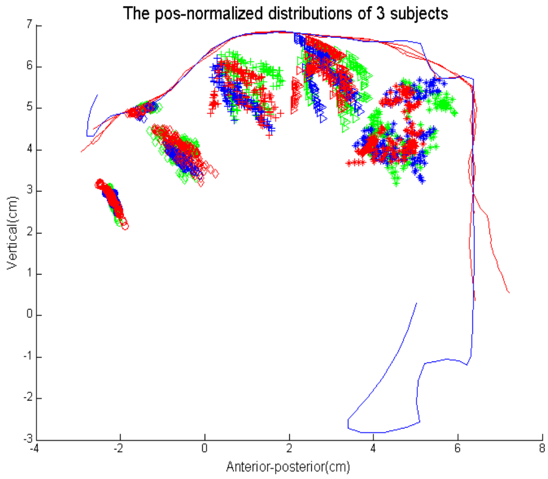

Figure 8. The transformed distributions of 8 consonants of three subjects are plotted. The symbols have same meaning as figure 7.

5 Evaluations

If the morphological variation reduced, the variance of articulatory data across subjects should be reduced. Cross-subject standard deviations (SD) of raw and normalized data of each tongue pellets are compared in Figure 9. There were 320 VCV tokens adopted in this research. For each token, the first vowel and central consonants were taken into account. In Figure 9, most data points fall below the line y=x, which indicates that the normalization reduced the cross-subject variances. In Table 1, the most normalized cross-subject standard deviations are comparable with the deviation of Subject TO, which indicates that after normalization the variances can be concentrated in a comparable range with a single subject.

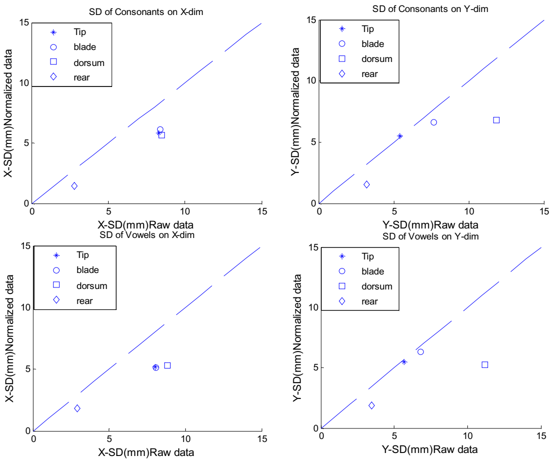

Figure 9. Comparisons of cross-subject standard deviations of raw and normalized coordinates of tongue pellets. The left panels show the x coordinates and right panels show the y coordinates. The upper panels show consonants' and lower panels show vowels'.

The crucial points are important for vocal tract to keep the posture for each phoneme during phonation. The Table 2 shows the differences between the distance from crucial points to palate of raw data and normalized data. The distance of crucial points to palate wall of vowels were changed less than 2 mm.

		Tip(x,y)		Blade(x,y)		Dorsum(x,y)		Rear(x,y)	
normalized all	Consonants	0.58	0.55	0.61	0.66	0.57	0.68	0.14	0.15
	Vowels	0.52	0.55	0.51	0.63	0.53	0.53	0.19	0.19
subject TO	Consonants	0.59	0.70	0.63	0.89	0.59	0.96	0.07	0.17
	Vowels	0.48	0.76	0.62	0.57	0.61	0.50	0.06	0.18

Table 1. Comparison of normalized SD of cross-subject with subject 2 (cm)

	Raw _dis	Normalized_dis	dis_diff	diff_perce
/a/ (dorsum)	1.47	1.64	-0.17	16.89%
/i/ (blade)	0.36	0.37	-0.01	1.25%
/u/ (dorsum)	0.86	0.72	0.14	13.95%
/e/ (dorsum)	1.06	0.87	0.19	18.58%
/o/ (dorsum)	1.29	1.42	-0.13	12.94%

Table 2. The distance from crucial points to palate of raw and normalized data and their differences (cm)

6 Conclusion

In this research, we proposed a mean to normalize the EMMA data across subjects by means of TPS transformation method. The performances of a normalization method of EMMA data were evaluated. The evaluation results showed that the inter-subject variations were reduced. The averaged standard deviations were reduced 2.1 mm for consonants and 2.3 mm for vowels over all tongue pellets.

Acknowledgments

The authors especially thank NTT Communication Laboratories for the permission to use the articulatory data. This study was supported in part by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan and in part by Grant-in-Aid for Scientific Research of Japan (No. 17300182).

References

[1] M. E. J. Beckman, T., T.-P. Jung, S.-h. Lee, K. d. Jong, A. K. Krishnamurthy, S. C. Ahalt, K. B. Cohen, and M. J. Collins, "Variability in the production of quantal vowels revisited," *J. Acoust. Soc. Am.,* vol. 97, pp. 471-490, 1995.

[2] M. Hashi, J. R. Westbury, and K. Honda, "Vowel posture normalization," *J. Acoust. Soc. Am.,* vol. 104, pp. 2426–2437, 1998.

[3] B. FL, "Principal warps: Thin plate splines and the decomposition of deformations," *IEEE Trans Pattern Anal. Mach. Intell,* vol. 11, pp. 567-85, 1989.

[4] T. Okadome and M. Honda, "Generation of articulatory movements by using a kinematic triphone model," *J. Acoust. Soc. Am,* pp. 453-463, 2001.

[5] L. Zagorchev and A. Goshtasby, " A comparative study of transformation functions for nonrigid image registration," *IEEE Trans. Image Processing,* vol. 15, pp. 529-538, 2006.

[6] J. Lim and M. H. Yang, "A Direct Method for modeling Non-rigid Motion with Thin Plate Spline," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[7] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *J. Acoust. Soc. Am.,* vol. 115, pp. 853-870, 2004.